# Programme

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00-8.20</td>
<td>Last week’s quiz results + Summary</td>
</tr>
<tr>
<td>8.20-9.00</td>
<td>Homology modelling – Part I</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.30</td>
<td>Exercise: Homology modelling</td>
</tr>
<tr>
<td>11.30-11.40</td>
<td>Break</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>

# Feedback Persons

![Feedback Persons Image]
Summary of Last Week

• Prediction of structural features from sequence.
  – Secondary structure
  – Disorder
  – Surface accessibility
  – Site predictions
    • Glycosylation, phosphorylation, signal peptides, TM helices…

Site predictions

• Phosphorylation
  – S/T or Y

• Glycosylation
  – N: N
  – O: S/T

\[
\text{NX}_T^S
\]
Programme

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00-8.20</td>
<td>Last week’s quiz results + Summary</td>
</tr>
<tr>
<td>8.20-9.00</td>
<td>Homology modelling – Part I</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.30</td>
<td>Exercise: Homology modelling</td>
</tr>
<tr>
<td>11.30-11.40</td>
<td>Break</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>

Homology Modelling
No Structure = No Go?

• Do homology modelling.
  – Use known protein structures to make reliable models.

• Validate the model.

• Adjust expectations and use accordingly!

Why Do We Need Homology Modelling?

• *Ab Initio* protein folding (random sampling):
  – 100 aa, 3 conf./residue gives approximately $10^{48}$ different overall conformations!

• Random sampling is *NOT feasible*, even if conformations can be sampled at picosecond ($10^{-12}$ sec) rates.
  – Levinthal’s paradox

• Do homology modelling instead.
How Is It Possible?

• The structure of a protein is uniquely determined by its amino acid sequence (but sequence is sometimes not enough):
  – prions
  – pH, ions, cofactors, chaperones

• Structure is conserved much longer than sequence in evolution.
  – Structure > Function > Sequence

How Often Can We Do It?

• Currently ~100,000 structures in the PDB
  – Approx. ~12000 with sequence ident. <30 % and resolution <3.0 Å.

• 25% of all sequences can be modelled.
• 50% can be assigned to a fold class.
Worldwide Structural Genomics

- "Fold space coverage"
- Complete genomes
  - Disease-causing organisms
  - Model organisms
- Membrane proteins
- Protein-ligand interactions

Hou et al., PNAS 2003, 100: 2386-2390

Homology Modeling & Structural Genomics

<table>
<thead>
<tr>
<th>Experimental Structure</th>
<th>Models or fold assignments</th>
<th>Models</th>
<th>Useful models</th>
<th>Less accurate models</th>
<th>Field assignments only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>1039</td>
<td>603</td>
<td>172</td>
<td>431</td>
<td>336</td>
</tr>
</tbody>
</table>

A model is counted if it is at least 40% accurate and is assumed to have ≤50% of its residues within 5 Å of their true position. "Models" are models with ≥80% sequence identity to the known structure, while "Less accurate models" are based on ≤50% sequence identity. "Field assignments only" denotes the number of predictions with a significant Fold recognition (FR) relationship to a known structure (60% identity) that failed to produce a reliable model. The calculations were performed in August 2006.

Protein Folds in PDB

How Well Can We Do It?

How Is It Done?

• Identify template(s)
  – Initial alignment
• Improve alignment
• Backbone generation
• Loop modelling
• Side chains
• Refinement
• Validation

Template Identification

• Search with sequence
  – Blast
  – Psi-Blast
  – Fold recognition methods
• Use biological information
• Functional annotation in databases
• Active site/motifs
Figure 25.3: A typical residue exchange or scoring matrix used by alignment algorithms. Because the score for aligning residues A and B is normally the same as for B and A, this matrix is symmetric.
Improving the Alignment

From "Professional Gambling" by Gert Vriend
http://www.cmbi.kun.nl/gv/articles/text/gambling.html
Template Quality

- Selecting the best template is crucial!
- The best template may not be the one with the highest % id (best p-value...)
  - Template 1: 93% id, 3.5 Å resolution
  - Template 2: 90% id, 1.5 Å resolution

The Importance of Resolution

- 4 Å resolution: low
- 3 Å resolution
- 2 Å resolution
- 1 Å resolution: high
- 0.5 Å resolution
Ramachandran Plot

- Allowed backbone torsion angles in proteins

Template Quality – Ramachandran Plot

X-ray structure – good data.  
NMR structure – low quality data…
Backbone Generation

- Generate the backbone coordinates from the template for the aligned regions.

- Several programs can do this, most of the groups at CASP6 use Modeller:
  
  http://www.salilab.org/modeller/

Loop Modelling – Connecting the Bits

- Knowledge based:
  - PDB search for fragments matching the target sequence (Levitt, Holm, Baker etc.).

- Energy based:
  - Uses an energy function to evaluate the quality of the loop and minimizes this function by Monte Carlo (sampling) or molecular dynamics (MD) techniques.
Loops – the Rosetta Method

• Find fragments (10 per amino acid) with the same sequence and secondary structure profile as the query sequence.

• Combine them using a Monte Carlo scheme to build the loop.

David Baker et al.

Side Chains

• Side chain rotamers are dependent on backbone conformation.

• Most successful method in CASP6 was SCWRL by Dunbrack et al.:  
  – Graph-theory knowledge based method to solve the combinatorial problem of side chain modelling.

http://dunbrack.fccc.edu/scwrl4/index.php
Side Chains

• Prediction accuracy:
  • High for buried residues
  • Low for surface residues
    – Experimental reasons: side chains at the surface are more flexible.
    – Theoretical reasons: much easier to handle hydrophobic packing in the core than the electrostatic interactions including H-bonds to waters.

Side Chains

• If the seq. id is high, the networks of side chain contacts may be conserved.
• Keeping the side chain rotamers from the template may be better than predicting new ones.
Refinement

- Energy minimization

- Molecular dynamics
  - Removes big errors
  - Atom clashes
  - Introduces small errors

- Keep minimization to a minimum.

Error Recovery

- Errors in the model can NOT be recovered at a later step
  - The alignment can not make up for a bad choice of template.
  - Loop modeling can not make up for a poor alignment.

- The step where the errors were introduced should be redone.
Validation

• Most programs will get the bond lengths and angles right.

• Model Rama. plot ~ template Rama. plot.
  – select a high quality template!

• Inside/outside distributions of polar and apolar residues.

Model Validation – ProQ

• ProQ is a neural network-based predictor
  – Structural features → quality of a protein model.

• ProQ is optimized to find
  – correct models…
  – …NOT (necessarily) native structures.

• Two quality measures:
  – MaxSub & LGscore

Arne Elofssons group: http://www.sbc.su.se/~bjorn/ProQ/
• LG score
  – Predict a structural alignment score:
  \[ S_{str} = M \left( \sum \frac{1}{1 + \left( \frac{d_{i,j}}{d_0} \right)} - \frac{N_{gap}}{2} \right) \]
  – Estimate the p-value (significance) of getting \( S_{str} \) compared to a random distribution.
  – Report LGscore as:
  \[ LGscore = -\log(p) \]

• MaxSub
  – % of residues in model predicted to be within 3.5 Å of the true position in a superimposition with the true structure.

LGscore vs. MaxSub

<table>
<thead>
<tr>
<th>Correct</th>
<th>Good</th>
<th>Very good</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGscore &gt; 1.5</td>
<td>LGscore &gt; 3</td>
<td>LGscore &gt; 5</td>
</tr>
<tr>
<td>MaxSub &gt; 0.1</td>
<td>MaxSub &gt; 0.5</td>
<td>MaxSub &gt; 0.8</td>
</tr>
</tbody>
</table>
Structure Validation Program Suites

- ProCheck
  [http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/](http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/)

- WhatIf server
  [http://swift.cmbi.ru.nl/whatif/](http://swift.cmbi.ru.nl/whatif/)

Homology Modelling Servers

- Eva-CM performs continuous and automated analysis of comparative protein structure modeling servers

- Current list of the best performing servers:
  [http://www.pdg.cnb.uam.es/eva/doc/intro_cm.html](http://www.pdg.cnb.uam.es/eva/doc/intro_cm.html)
Summary

• Successful homology modelling depends on the following:
  – Template quality
  – Alignment (add biological information)
  – Modelling program/procedure (use more than one)

• Always validate your final model!

Programme

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00-8.20</td>
<td>Last week’s quiz results + Summary</td>
</tr>
<tr>
<td>8.20-9.00</td>
<td>Homology modelling – Part I</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.30</td>
<td>Exercise: Homology modelling</td>
</tr>
<tr>
<td>11.30-11.40</td>
<td>Break</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>
## Programme

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00-8.20</td>
<td>Last week’s quiz results + Summary</td>
</tr>
<tr>
<td>8.20-9.00</td>
<td>Homology modelling – Part I</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.30</td>
<td>Exercise: Homology modelling</td>
</tr>
<tr>
<td>11.30-11.40</td>
<td>Break</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>